COPY 1 of 1 COPIES

CONCERNING THE EFFECT OF SMALL CORRELATION ON CERTAIN LARGE SAMPLE TESTS AND CONFIDENCE INTERVALS FOR THE MEAN

John E. Walsh

P-112

17 November 1949

CONCERNING THE EFFECT OF SMALL CORRELATION ON CERTAIN LARGE SAMPLE TESTS AND CONFIDENCE INTERVALS FOR THE MEAN

By John E. Walsh
The RAND Corporation

Summary. Most of the well known significance tests and confidence intervals for the population mean are based on the assumption of a random sample. This paper considers how the significance levels and confidence coefficients of a commonly used class of these tests and intervals are changed when the random sample requirement is violated and the number of observations is large. It is found that the introduction of even a slight amount of correlation can result in a substantial significance level and confidence coefficient change. Thus this class of tests and confidence intervals would seem to be of questionable practical value for large sets of observations. For two types of situations of practical interest, methods are outlined for obtaining large sample tests and confidence intervals for the mean which are not sensitive to the presence of correlation. These results are as efficient (asymptotically) as the tests and intervals they replace and are applicable to the general situation where the observations are not from the same population.

Introduction and Statement of Results. In deriving statistical tests and confidence intervals, certain assumptions are made. When these tests and intervals are applied to practical situations, their validity depends upon how closely the assumptions are approximated. One assumption frequently made is that a set of observations is a random sample; i.e., that the observations are

(a). Statistically independent.

(b). From the same population (i.e., the observations have the same univariate distribution function).

One of the principal purposes of this paper is to study how sensitive a commonly applied class of large sample tests and confidence intervals for the population mean is to violation of assumption (a).

Let the n observations used be from populations with common mean μ while the values of the observations are denoted by $x_1, \cdots, x_n$. The class of tests and confidence intervals for μ investigated here consists of those based on the quantity

$$(1) \qquad \sqrt{n}\,(\bar{x} - \mu) \Big/ \sqrt{\sum_1^n (x_i - \bar{x})^2/(n-1)} \quad ,$$

where $\bar{x} = \sum x_i/n$. If the n observations are a random sample from a population for which the first two moments exist (almost all populations approximated in practice have this property), the distribution of (1) is approximately normal with zero mean and unit variance for n sufficiently large. The quantity (1) is the well known Student t-statistic; if the observations were a random sample from a normal population, this quantity would have a Student t-distribution with $n - 1$ degrees of freedom ($n \geq 2$).

For large values of n, it is found that the introduction of an average correlation as small as .001, or even .0001, can result in substantial changes in the values of significance levels and confidence coefficients for the class of tests and intervals based on (1). As

average correlations of this magnitude usually defy detection, it is possible that such correlations exist in many practical situations where a random sample is assumed. Thus it would seem important to replace this class by other tests and confidence intervals which are not sensitive in this respect. Replacement results are derived for two special types of situations:

(i). The observations can be divided into two or more nearly equal subsets such that the correlation within subsets is very much greater (in magnitude) than the correlation among subsets.

(ii). The order in which the observations were drawn is known. The correlation between nearby observations in this ordering is very much greater (in magnitude) than the correlation between distant observations.

As an example of (i), consider an agricultural experiment where part of the observations come from one locality, part from another locality, etc. For this case it is often permissible to assume approximate independence between observations from different localities and noticeable correlation between observations from the same locality. If the localities do not yield approximately the same number of observations, subsets satisfying (i) can frequently be obtained by combining localities.

Now let us consider situation (ii). If the observations are obtained from a process which produces one observation at a time, in many cases it is reasonable to assume that the correlation between two observations depends on their location in the time ordered sequence of observations. Two observations distant from each other are almost independent while observations near each other are appreciably correlated.

Consider case (i) and let the n observations be divided into r approximately independent subsets. The subsets contain about the same number of observations and the s_k observations of the k^{th} subset are given the special notation

$$x_1(k), x_2(k), \cdots, x_{s_k}(k), \qquad (k = 1, \cdots, r).$$

Actually, this set of observations is a subset of the observations $x_1, \cdots, x_n$ but was given an additional notational representation to simplify presentation of results. Thus, for this case, each observation has two representations, one as an x_i and the other as an $x_j(k)$. This double representation should result in no confusion since the only places in which the x_i notation is used are those where all n observations are considered simultaneously and given identical treatment (e.g., $\sum_1^n x_i^2$, etc.). The tests and confidence intervals developed for μ are based on the quantity

$$(2) \qquad \sqrt{n}\,(\bar{x} - \mu) \Big/ \sqrt{\left| \frac{1}{n}\sum_1^n x_i^2 + A \sum_{k=1}^{r} \sum_{u \neq v=1}^{s_k} x_u(k)x_v(k) + B \sum_{i \neq j=1}^{n} x_i x_j \right|}\,,$$

where

$$A = n \Big/ \left[n(n-1) - \sum_{k=1}^{r} s_k(s_k - 1) \right], \qquad B = 1/n - A.$$

If n is large and condition (i) holds, the distribution of (2) is approximately normal with

zero mean and variance approximately unity for most situations of practical interest.

Under some restrictions which are not of practical importance, it can be shown that the asymptotic distribution of (2) is normal with zero mean and unit variance when the n observations are a random sample. Thus the results based on (2) have the same efficiency (asymptotically) as the corresponding tests and confidence intervals based on (1) for the case of a random sample. However, the results based on (2) are also valid for case (i).

Now let us consider case (ii). Then the order in which the observations were drawn is known. Let x_1 denote the value of the first observation drawn, x_2 the value of the second observation, $\cdots$, x_n the value of the last observation drawn. It is supposed that an integer value m is known ($m < n/2 - 1$) such that the correlation between any two observations x_i, x_j can be neglected if $|i - j| > m$. Then the tests and confidence intervals derived for μ are based on the quantity

$$(3) \qquad \sqrt{n}\,(\bar{x} - \mu) \Big/ \sqrt{\left| \frac{1}{n} \sum_1^n x_i^2 + C \sum_1^n x_i (x_{i+1} + \cdots + x_{i+m}) + D \sum_{i \neq j = 1}^{n} x_i x_j \right|} \quad ,$$

where

$$C = 2/(n - 2m - 1), \qquad D = -(2m + 1)/n(n - 2m - 1).$$

In this expression, x_i's appear for which $i > n$; by definition, $x_{n+j} = x_j$ for all such cases. If n is sufficiently large and (ii) is satisfied, the distribution of (3) is nearly normal with zero mean and variance approximately equal to unity for most situations of practical importance.

For the case of a random sample, under mild restrictions it can be shown that the asymptotic distribution of (3) is normal with zero mean and unit variance. Hence the results based on (3) furnish the same amount of "information" (asymptotically) as the corresponding results based on (1) for the case of a random sample.

In general, the value chosen for m should not be too small; otherwise the neglected effect of a large number of small correlations might become important. On the other hand, too large a value of m results in computational and other difficulties for the solution presented. A rule of thumb which is likely suitable for most situations is to set $m \triangleq n/100$ for situations where there is reason to believe that smaller values of m would be satisfactory.

What is meant by the statement "a sufficiently large value of n" as used in this section is difficult to specify in general. However, $n \geq 1{,}000$ would seem to be satisfactory for many situations. The results presented apply to both discrete and continuous variables.

Investigation of Specified Class. Now let us investigate how sensitive the tests and confidence intervals based on (1) and the assumption of a random sample are to slight violation of condition (a). Denote the variance of x_i by σ_i^2 and the correlation between x_i and x_j by ρ_{ij}, $(i \neq j = 1, \cdots, n)$. Also let each observation have the same expected value μ.

Imposing some weak restrictions on the behavior (as n increases) of the fourth and lower order moments (mixed or otherwise) of the multivariate population from which $x_1, \cdots, x_n$

is a sample value, the variance of

$$\sum_{1}^{n} (x_i - \bar{x})^2/(n - 1)$$

tends to zero as n approaches infinity. Thus, since the expected value of this expression equals

$$\frac{1}{n}\sum_{1}^{n} \sigma_i^2 - \frac{1}{n}\sum_{i\neq j=1}^{n} \rho_{ij}\sigma_i\sigma_j/(n - 1),$$

it follows from Tchebycheff's Inequality combined with the convergence theorem [1] that asymptotically the distribution of

$$(4) \qquad \left[\frac{\sum \sigma_i^2 - \sum_{i\neq j} \rho_{ij}\sigma_i\sigma_j/(n - 1)}{\sum \sigma_i^2 + \sum_{i\neq j} \rho_{ij}\sigma_i\sigma_j}\right]^{\frac{1}{2}} \frac{\sqrt{n}\,(\bar{x} - \mu)}{\sqrt{\sum (x_i - \bar{x})^2/(n - 1)}}$$

has zero mean and unit variance. Hence, for large n, the variance of (1) differs from its hypothetical value of unity by the factor

$$\left[1 + (n - 1)\varphi\right]/(1 - \varphi),$$

where

$$\varphi = \sum_{i\neq j} \rho_{ij}\sigma_i\sigma_j/(n - 1)\sum \sigma_i^2 .$$

The range of permissible values for φ is $-1/(n - 1)$ to 1. If the σ_i have the same value, φ represents the average correlation among the observations; i.e.,

$$\varphi = \sum_{i\neq j} \rho_{ij}/n(n - 1).$$

For most situations of practical interest, the distribution of

$$\sqrt{n}\,(\bar{x} - \mu)\Big/\sqrt{\frac{1}{n}\left(\sum \sigma_i^2 + \sum_{i\neq j} \rho_{ij}\sigma_i\sigma_j\right)}$$

is asymptotically normal. Thus, using Tchebycheff's Inequality and the convergence theorem [1], for these situations the distribution of (4) is approximately normal with zero mean and unit variance for n sufficiently large. Let us consider a class of one-sided confidence intervals for μ obtained by use of (1) under the assumption of a random sample; here it is assumed that the distribution of (4) is approximately normal. Let K_ϵ denote the standardized normal deviate (zero mean, unit variance) exceeded with probability ϵ. The one-sided confidence interval

$$(5) \qquad \left(\bar{x} + K_\epsilon \sqrt{\sum (x_i - \bar{x})^2/(n - 1)}\ ,\ \infty\right)$$

is then assumed to have confidence coefficient ϵ; i.e.,

$$\Pr\left(\bar{x} + K_\epsilon \sqrt{\sum (x_i - \bar{x})^2/(n - 1)} < \mu\right) = \epsilon.$$

Actually,

$$\Pr\left[\bar{x} + K_\epsilon \sqrt{\sum (x_i - \bar{x})^2/(n-1)} < \mu\right]$$

$$= \Pr\left\{\bar{x} + K_\alpha \sqrt{[1 + (n-1)\rho]\sum (x_i - \bar{x})^2/n(n-1)(1-\rho)} < \mu\right\}$$

$$= \alpha,$$

where α is defined by the relation

$$K_\alpha = K_\epsilon \sqrt{(1-\rho)/[1 + (n-1)\rho]}\ .$$

If the observations are a random sample, $\rho = 0$ and $\alpha = \epsilon$. If $\rho \neq 0$, however, α can differ noticeably from ϵ. For example, let $\epsilon = .05$, $n = 10{,}000$ and $\rho \geq .001$. Then $\alpha \geq .31$. If $\rho \leq -.00008$ for this case, $\alpha \leq .00012$. Thus very slight deviations of ρ from zero can result in substantial deviations of the true confidence coefficient of (5) from its hypothetical value. Analogous considerations apply to other one-sided and two-sided confidence intervals and to significance tests based on these confidence intervals.

From the definition of ρ and the analysis of this section, it is seen that only slight average correlations need be introduced to cause the significance levels and confidence coefficients of tests and intervals based on (1) to differ substantially from their assumed values. Thus results based on (1) and the assumption of a random sample are very sensitive to slight deviations from (a).

Derivations for Case (i). Here it will be shown that the asymptotic distribution of (2) is normal with zero mean and variance approximately equal to unity if (i) and certain minor restrictions are satisfied.

For the purpose of the analysis, it will be assumed that $s_1 = s_2 = \cdots = s_r$. The results obtained on the basis of this assumption should not be appreciably changed if these equalities are only approximately satisfied. Using the restriction on the s_k, the expression

$$(6) \qquad \frac{1}{n}\sum_1^n x_i^2 + A\sum_{k=1}^{r}\sum_{u\neq v=1}^{s_k} x_u(k)x_v(k) + B\sum_{i\neq j=1}^{n} x_i x_j$$

can be written as a linear combination of the quantities

$$\sum_1^n (x_i - \bar{x})^2, \qquad \sum_{u=1}^{s_k}[x_u(k) - \bar{x}(k)]^2, \qquad (k = 1, \cdots, r),$$

where

$$\bar{x}(k) = \sum_1^{s_k} x_u(k)/s_k.$$

Thus replacing each x by $x - \mu$ leaves the value of (6) unchanged.

Consider the expected value of (6). Subtracting μ from each x in (6) and then taking expected values, this is found to equal

$$\frac{1}{n}\sum_1^n \sigma_i^2 + AE\left\{\sum_k \sum_{u\neq v} [x_u(k) - \mu][x_v(k) - \mu]\right\} + B\sum_{i\neq j} \rho_{ij}\sigma_i\sigma_j.$$

On the basis of condition (i),

$$E\left\{\sum_{k}\sum_{u\neq v}[x_u(k)-\mu][x_v(k)-\mu]\right\} \doteq \sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j$$

for reasonable types of situations. Thus the expected value of (6) is approximately equal to

$$\frac{1}{n}\sum_{1}^{n}\sigma_i^2 + \frac{1}{n}\sum_{i\neq j=1}^{n}\rho_{ij}\sigma_i\sigma_j. \tag{7}$$

With some mild restrictions on how the fourth and lower order moments of the multivariate population from which $x_1, \cdots, x_n$ was drawn behave as n increases, the variance of (6) tends to zero as n approaches infinity. Since the expected value of (6) is near the value of (7), this expected value is positive for almost any situation of practical interest. Consequently, it follows from Tchebycheff's Inequality that $\Pr[(6) < 0] \longrightarrow 0$ as $n \longrightarrow \infty$. From this it is seen that the expected value of the absolute value of (6) is also approximately equal to (7) for large n.

As the variance of $\sqrt{n}\,(\bar{x} - \mu)$ has the value (7), the properties stated for (2) follow from Tchebycheff's Inequality combined with the convergence theorem [1].

Derivations for Case (ii). In this section it will be shown that the asymptotic distribution of (3) is normal with zero mean and variance approximately equal to unity if (ii) and certain minor restrictions hold.

The expression

$$\frac{1}{n}\sum_{1}^{n}x_i^2 + C\sum_{1}^{n}x_i(x_{i+1} + \cdots + x_{i+m}) + D\sum_{i\neq j=1}^{n}x_ix_j \tag{8}$$

can be written as a linear combination of the quantities

$$\sum_{1}^{n}(x_i - \bar{x})^2, \qquad \sum_{1}^{n}\left[(x_i - x_{i+1})^2 + \cdots + (x_i - x_{i+m})^2\right].$$

Consequently replacing each x by $x - \mu$ leaves the value of (8) unchanged.

Let us consider the expected value of (8). Replacing each x in (8) by $x - \mu$, this expected value is found to be approximately equal to (7) if

$$E\left\{2\sum_{1}^{n}(x_i - \mu)\left[(x_{i+1} - \mu) + \cdots + (x_{i+m} - \mu)\right]\right\} \doteq \sum_{i\neq j=1}^{n}\rho_{ij}\sigma_i\sigma_j. \tag{9}$$

On the basis of condition (ii) and the relation $m \geq n/100$, it appears that (9) will be satisfied for most situations of practical importance.

Imposing some weak restrictions on the behavior (as n increases) of the fourth and lower order moments of the multivariate population from which $x_1, \cdots, x_n$ was obtained, the variance of (8) tends to zero as n tends to infinity. Since the expected value of (8) is near (7), this expected value will almost always be positive for situations approximated in practice. Hence, from Tchebycheff's Inequality, $\Pr[(8) < 0] \longrightarrow 0$ as $n \longrightarrow \infty$. Thus, asymptotically the expected value of $|(8)|$ is also approximately equal to (7).

The properties stated for (3) now follow from a combination of Tchebycheff's Inequality and the convergence theorem [1].

REFERENCE

[1] Harald Cramér, Mathematical Methods of Statistics, Princeton Univ. Press, 1946, p. 254.